HP P6300/P6500 Enterprise Virtual Array Best Practice

Technical white paper

Table of contents

Abstract

An important value of the HP P6000 Enterprise Virtual Array (EVA) is simplified management. A storage system that is simple to administer saves management time and money, and reduces configuration errors. You can further reduce errors and unnecessary expense by implementing a few best practices and optimizing your P6000 EVA for their intended applications. This paper highlights common configuration rules and tradeoffs for optimizing the new HP P6300 and P6500 Enterprise Virtual Array for cost, availability, and performance. Getting the most from your enterprise-class storage has never been easier.

Background

Two design objectives for the HP P6000 EVA are to provide maximum real-world performance and to reduce storage management costs. These objectives result in the designing of an intelligent controller that reduces the number of tuning parameters that are user controlled. In contrast, traditional disk arrays typically have many tunable settings for both individual logical unit numbers (LUNs) and the controller. Although tunable settings might appear to be desirable, they pose potential issues:

- It is difficult and time consuming for administrators to set the parameters appropriately. Many settings require in-depth knowledge of the controller's internal algorithms and specific knowledge of the workload presented to the array.
- Storage administrators often do not have the time and resources to attain the expertise necessary to maintain a traditional array in optimum configuration.
- As the I/O workload changes, many parameters that were previously set might no longer be appropriate. Overcoming the effects of change requires continual monitoring, which is impractical and costly in most situations.

Because of such concerns, HP P6000 EVA algorithms reduce the parameters that users can set, opting instead to embed intelligence within the controller. The controller has a better view of the workload than most administrators and can be far more dynamic in responding to workload changes. The result is an array that is both easy to configure and high performing.

Overview

Although the HP P6000 EVA is designed to work in a wide range of configurations, configuration options can influence performance, usable capacity, or availability. With the necessary information about the configuration options for the HP P6000 EVA, a storage administrator can enhance its configuration for a specific application.

These configuration choices include:

- Number of disks
- Number of disk groups
- Type and number of disks in a disk group
- Vraid levels (0, 1, 5, and 6)
- Disk failure protection level (none, single, and double)
- Disaster Recovery (DR) groups, snapshots, snapclones, and mirrorclones (In this paper, the term "clone" is used where appropriate to refer collectively to both snapclones and mirrorclones.)
- Application configuration and third-party disaster-recovery solutions
- Cache settings
- Capacity management
- Thin Provisioning
- Dynamic LUN/RAID migration

Each of these topics is detailed in the following sections. Note that it may not be possible to enhance a configuration for cost, performance, and availability, simultaneously. Sometimes conflicting recommendations defeat the purpose of one objective while trying to accommodate the demands made by other objectives. For example, Vraid0 is clearly the best solution from a strict cost standpoint because nearly all storage is available for user data. However, Vraid0 offers no protection from a disk failure; thus, Vraid1, Vraid5, or Vraid6 is a better choice for availability. Other tradeoffs can be more complex by comparison but are worth understanding because there is no best choice in all situations. "Best" depends on the priorities of a particular environment.

The best practices in this paper are based on controller software version 10.00.00.00 and later.

Best practices summary

The following table summarizes typical best practices for optimum availability, performance, and cost. As with most generalizations, they are not the best choice for all applications. For detailed information, see the associated sections of this paper. As best practices, they are recommendations, not requirements. The HP P6000 EVA supports a wide variety of configurations, and all supported configurations provide availability and performance features. These best practices have been developed to help you make good configuration choices when alternatives exist. For configuration requirements, see the P6000 EVA user manual or installation guide.

Table 1: Best practices summary

Best practice	Source	Discussion
Disks in disk groups in multiples of eight	Availability	Enable the P6000 EVA to optimize the distribution of disks in the Redundancy Storage Set (RSS).
Use disks of same size and speed in a disk group	Performance	Improves ease of management and cost utilization. This configuration avoids any issues with access-density of the disk group.
As few disk groups as possible	Performance, Cost	Performance is enhanced when the array is allowed to stripe data to as many disks as possible.
Protection level of one	Cost	In most installations, a protection level of one provides adequate availability. See detailed discussion of mitigating conditions.
Separate disk group for database logs	Availability	Provides consistent and current database restore from external media if data/table space disk group is inconsistent.
Use Solid State disks	Performance	Solid State disks have the highest performance. However, they are the most expensive. 15k rpm disks have equal or higher performance than 10k rpm disks, but they are more expensive. See details for discussion on price-performance optimization.
Load balance demand to controllers	Performance	Balancing the workload as evenly as possible to both controllers provides the highest performance utilization.
Vraid1	Availability, Performance	Provides the best combination of performance, data protection, and availability. For most workloads, Vraid1 provides the best performance.
Vraid5	Cost	Provides the lowest cost of protected storage.
Vraid6	Availability	Provides the highest level of data protection and availability.
Capacity management	Availability	Proper settings for the protection level, occupancy alarm, and available free space provide the resources for the array to respond to capacity-related faults.
HP P6000 Continuous Access or host-based mirroring	Availability	Real-time mirroring to an independent array provides the highest levels of data protection and availability. Geographically dispersed arrays provide disaster protection.
External media backup	Availability	All data center best practices include processes to regularly copy the current data set to external media or near-line devices.
Insight Remote Support	Availability	Provides automated messaging of abnormal EVA status to HP Support or your internal IT personnel.

First best practices

The first best practices are generally common sense.

- Read the P6000 EVA user manual. Always operate the array in accordance with the user manual. In particular, never exceed the environmental operation requirements.
- Use the latest controller and disk firmware to benefit from the continual improvements in the performance, reliability, and functionality of the P6000 EVA. For additional information, see the release notes and release advisories for the respective EVA products.
- Deploy the array only in supported configurations. In many cases, HP does not support a particular configuration if it failed our testing. Do not risk the availability of your critical applications to unsupported configurations.

- Sign up for proactive notifications at: http://www.hp.com/go/myadvisory. Receiving update notifications and applying the suggested resolutions will enhance availability.
  - First best practice: Read and adhere to the user manual.
  - First best practice: Stay as current as possible with the latest XCS controller software and disk firmware.
  - First best practice: Deploy the array in supported configurations only.
  - First best practice: Sign up for proactive notifications.

Best practices to reduce cost

The term "cost" in this paper refers to the cost per unit of storage. The cost is obtained by dividing the total cost of the storage by the usable data space as seen by the operating system. Cost is typically expressed in dollars per MB or dollars per GB. This section discusses options for improving the total usable capacity for a given configuration, thus lowering the cost per unit of storage.

Mixed disk capacities influence the cost of storage

The HP P6000 EVAs can simultaneously support disks of several different capacities. Larger disks are usually more expensive but offer a lower price per unit of storage. In addition, disk technology has historically doubled capacity points every 18 months. The result of these market and technical factors is that storage arrays tend to be configured, or at least there is interest in configuring them, with disks of different capacities. The P6000 EVA greatly simplifies the management of these configurations. However, understanding configuration guidelines can help you improve a solution using disks with different capacities.

The disk-failure protection level is a selectable parameter, which defines the number of disk failure and auto reconstruction cycles a disk group can tolerate before failed disks are replaced. The protection level can be dynamically assigned to each disk group as a value of none, single, or double. Conceptually, it reserves space to handle 0 (none), 1 (single), or 2 (double) disk failure-reconstruct cycles. The space reserved is specific to a particular disk group and cannot span disk group boundaries.

The software algorithm for reserving reconstruction space finds the largest disk in the disk group; doubles its capacity; multiplies the result by 0, 1, or 2 (the selected protection level); and then removes that capacity from free space. Unlike traditional arrays, the P6000 EVA does not reserve physical disks. The reconstruction space is distributed across all disks in the disk group so that all disks remain available for application use. This is called distributed sparing. The largest disk is used even though there might only be a few of them in a disk group. By using the capacity of the largest disk for spare space, all disks in the disk group are protected, from the smallest capacity disks up to and including the largest capacity disks.

The reason the algorithm doubles the disk count is that the Vraid1 recovery algorithm requires Vraid1 spare space to be in predetermined disk pairs. When a member of a pair fails, the remaining contents are moved to a new pair; so twice the capacity is reserved. The advantages of distributed sparing are two-fold. First, the performance value of all disks is used, and second, unlike a traditional spare, there is no risk that a reserved disk is unusable (failed) when needed.

As reconstruction space is not shared across disk groups, it is more efficient to have the least number of disk groups possible, thus reducing the reconstruction overhead. When the first few larger disks are introduced into a disk group, the resulting usable capacity is similar to the smaller disks until the protection-level requirements are met. Then, the additional usable capacity is in line with the physical capacity of the new disks. Even though the resulting capacity for the first few disks is not very efficient, the alternative of creating two disk groups provides less usable capacity.

- **Best practice to reduce the cost of storage: Mix disk sizes within a single disk group.**

Number of disk groups influences the cost of storage

Disk groups are independent protection domains. All data redundancy information and reconstruction space must be contained within the disk group. Unlike traditional RAID 5 arrays, where the stripe depth and redundancy set are the same, the P6000 EVA supports many disks in the stripe set (a disk group), but the redundancy set (RSS) is always six to 11 disks. See Best practices to improve availability. As a result, multiple disk groups are not needed for Vraid5 availability and would provide excessive reconstruction space for the small number of disks that would be in each disk group.

As with traditional arrays, multiple disk groups can result in stranded capacity. Stranded capacity occurs when the capacity that could be used to create a Vdisk is distributed to many disk groups, and so, cannot be used to create a single LUN. (A Vdisk must be created within a single disk group.) Unlike traditional disk arrays, the P6000 EVA can easily support very large disk groups, eliminating stranded capacity issues.

- **Best practice to reduce the cost of storage: Create as few disks groups as possible.**

Note:
To understand the tradeoffs associated with the number of disk groups, see the discussion of disk groups in Best practices to improve availability and Best practices to enhance performance. Before acting on any best practice, you should understand the effect on price, performance, and availability.

Number of disks influences the cost of storage

The lowest cost per storage unit is achieved by allocating the cost of the controllers over as many disks as possible. The lowest cost per storage unit for disks is typically on the largest disks. Thus, the lowest-cost solution is to fill the array with as many of the largest disks as are supported. This configuration results in the lowest cost of storage per MB.

- **Best practice to reduce the cost of storage: Fill the P6000 EVA with as many disks as possible, using the largest-capacity disks.**

Note:
Before increasing the disk count, see Best practices to improve availability and Best practices to enhance performance.

Thin Provisioning influences the cost of storage

Thin Provisioning software allows customers to present applications with more capacity than is physically allocated to them in the array. This helps reduce physical capacity requirements and therefore lowers cost. As an application writes data to the thin provisioned Vdisk, the controller firmware will automatically allocate more space, to match the size of the Vdisk.

Thin provisioning allows disk purchases to be delayed. With fewer disks, power and cooling cost increases are also delayed. Managing the disk provisioning is simplified, due to two thresholds that can be set (in addition to the critical threshold notification). The first one has the ability to set capacity notification thresholds, where alerts are sent out when the threshold is reached, informing the user that more capacity must be added. The second one is the Vdisk capacity threshold, which indicates that an application is not thin-friendly and so automates allocation of the capacity for that Vdisk. These two new thresholds can simplify capacity management for the storage administrator since they won't have to actively monitor the capacity of the disk group or the Vdisk.

Thin provisioned Vdisk allocation: The P6000 EVA allocates virtual mapping capacity for a thin provisioned Vdisk although it does not allocate space on the disk group. Each array model has a different amount of mapping memory available for Vdisks, so it is important not to create larger thin provisioned Vdisks than what is needed. If required, the user can increase the size of a thin provisioned Vdisk online, without any application disruption. DCM software can also be used to automatically increase the Vdisk size as the allocated capacity increases.

- **Best practice to reduce the cost of storage: Size thin provisioned Vdisk for the expected capacity needed by the application over the next year or two, rather than by the maximum of 32 TB.**

Thin provisioned Vdisk alarm setting: P6000 EVA firmware XCS 10.00.00.00 provides an alarm that gets triggered when a thin provisioned Vdisk is allocating more space than expected. This alarm helps to identify runaway applications that may not be "thin provisioning friendly."

- **Best practice to reduce the cost of storage: Set the thin provisioned Vdisk alarm to just above the expected allocation for the Vdisk.**

HP Dynamic Capacity Management influences the cost of storage

HP Dynamic Capacity Management (DCM) enables the storage administrator to reduce the problem of over-provisioning storage to an application. To accomplish this, the administrator needs to prevent an application from running out of space, while at the same time minimizing the amount of wasted space dedicated to an application at any point in time.

DCM leverages and uses the capabilities of EVA virtualization to provide an efficient provisioning solution. DCM takes advantage of the unique EVA characteristics while using non disruptive LUN extend or shrink capabilities to make efficient provisioning a simple task for the storage administrator. The purpose of DCM is to allow a storage administrator to provision storage to an application by understanding its short-term storage needs, its growth rate, and its long-term storage needs.

For more information on DCM, including best practices, see the HP Enterprise Virtual Array Dynamic Capacity Management best practices white paper, which can be downloaded from the following HP website: http://h18006.www1.hp.com/storage/arraywhitepapers.html

Dynamic LUN/RAID Migration influences cost

Dynamic LUN and RAID Migration: The MIGRATE command changes the Vraid of a virtual disk, moves a synchronized standalone virtual disk from one disk group to another, or does both on the EVAx400 and P6000 arrays without disrupting host workload. This operation uses a mirrorclone to copy the data to the designated disk group and therefore requires a business copy license. Once the data in the mirrorclone has been synchronized with the source Vdisk, P6000 Command View will issue a command to exchange the roles of the source Vdisk and the mirrorclone. Host I/Os are automatically transferred to the new source Vdisk. Because a mirrorclone is used for the migration of data, additional capacity is temporarily needed until the migration has completed. An option in command view allows the user to keep or delete the resulting mirrorclone after the migration is completed. Use this option to reverse the changes if needed.

- **Best practice to reduce the cost of migration: Use the Dynamic LUN and RAID Migration feature present in XCS 10.00.00.00 to manage the Vdisk migration or change the raid level, efficiently, while the host I/O operations are still active.**

Disk performance influences the cost of storage

Larger disks usually offer better price-per-capacity than smaller disks. Although prices continuously change, more capacity can be purchased for the same price by purchasing larger drives. Conversely, higher performance drives, such as 15k rpm drives, are generally more expensive than their lower performance 10k rpm counterparts.

- **Best practice to reduce the cost of storage: Use lower performance, larger capacity disks.**

Cross-RAID snapshots influence the cost of storage (and the perceived availability of storage)

The P6000 EVA allows the target LUN of a demand-allocated or fully-allocated snapshot to be a different Vraid type than the source. For example, a Vraid6 LUN can have an associated Vraid0, Vraid1, Vraid5, or Vraid6 snapshot LUN. With demand-allocated snapshots, the capacity is allocated as writes are copied to the demand-allocated snapshot. This contrasts with fully allocated snapshots which, at time of creation, allocate physical space equivalent to the full capacity of the parent LUN. Only changed data sets are copied. This means that the unchanged data remains in the source LUN (and RAID level), and the changed data resides in the snapshot target LUN (and RAID level).

The availability characteristics of a cross-RAID snapshot LUN are those of the lower (in an availability sense) RAID level of either the source or the target LUN. Therefore, it does not make economic sense to assign the snapshot target LUN a RAID level with greater availability than the source. The snapshot LUN would consume more capacity without providing greater availability.

The hierarchy of RAID levels for the EVA is:

1. Vraid6—highest availability, uses some capacity for parity protection.
2. Vraid1—highest availability, uses the most raw capacity.
3. Vraid5—high availability, uses some capacity for parity protection.
4. Vraid0—lowest availability, uses the least raw capacity.

- **Best practice to reduce the cost of storage: Use an equal or lower RAID level for the target LUN of a Vraid snapshot.**

SAS Mid-line disks influence the cost of storage

SAS Mid-line disks are low-cost, low-performance disks for use in the P6000 EVA. Because of the design of these disks, HP recommends a reduced duty cycle to meet business-critical availability requirements.

Note:
Reduced duty cycle can be defined as the disk drive workload or duty cycle = (time reading or writing data) / the measurement time. For mid-line drives, the recommended duty cycle is less than or equal to 0.4.

Use mid-line disks only where random-access performance and continuous operation are not required. The EVA requires that mid-line disks be organized in disk groups that are separate from other drive types.

The best application for mid-line disks is for the online part of your backup and recovery solution. Clones assigned to mid-line disk groups provide the lowest-cost solution for zero-downtime backup and fast recovery storage. HP software solutions, such as HP Data Protector, and other industry backup software, include the management of mid-line based clones for backup and recovery.

Mid-line drives are not recommended in Continuous Access applications as the remote storage location for local data residing on standard higher speed disk drives. Continuous Access tends to perform high-duty cycle random writes to the remote disk array. Matching remote mid-line disks with local enterprise disks will affect the performance of your application, and will adversely impact the reliability of the mid-line disks.

- **Best practice for SAS Mid-line disk and the cost of storage: Use mid-line disks to augment near-line storage usage.**

Note:
SAS Mid-line disks and clone copies are not a replacement for offline backup. Best practice is to retain data on an external device or media for disaster recovery.

Best practices to improve availability

The P6000 EVA is designed for business-critical applications. Redundancy features enable the array to continue operation after a wide variety of failures. The P6000 is also designed to be flexible. Some configurations allow higher availability by limiting exposure to failures that exceed the fault-tolerant design of the array. The goal is to create configurations that have the most independent protection domains, so that a failure in one domain does not reduce the resiliency of another domain.

All supported configurations have some tolerance to failures, but not necessarily to all failures. As an example, Vraid0 offers no protection from a disk failure, but it is still resilient to most back-end Fibre Channel loop failures.

The following guidelines focus on two areas. They address configurations to improve availability in sequential and multiple simultaneous failure scenarios. And, they also discuss system and application configurations to improve availability.

Backup processes to enhance recovery

Independent of the online storage device, make sure you include a proven backup and recovery process in array and data center management procedures. The P6000 EVA is supported by numerous backup applications. HP Data Protector and similar third-party backup software are supported on a variety of popular operating systems. They support the EVA directly or are integrated through Oracle, Microsoft® SQL Server, or other databases to provide zero-downtime backup.

Along with the backup data, save a copy of the initial HP Command View related configuration files, as well as a copy of the files when a configuration change occurs. An exact description of the array configuration greatly reduces recovery time. The backup configuration files should be stored on external media—not on the array.

Do not consider array-based copies as the only form of backup. Snapshot and clone copies complement a backup strategy that includes full copies to offline or near-line storage. In this application, clones can provide alternatives for reducing recovery time by providing the first option for recovery.

Perform regular backups and be sure to test the restore process twice a year. The P6000 EVA greatly simplifies testing by providing a simple process to create and delete disk groups or Vdisks. Capacity used for testing can be easily reused for other purposes.

- **Best practice to maintain data recovery: Perform regular data backup and biannual recovery tests.**
- **Best practice to maintain data recovery: Include a copy of the P6000 EVA configuration with the backup data. This can be accomplished with the HP Storage System Scripting Utility Capture Configuration command.**

Disk group type influences availability

The current disk group type will store metadata in a double Vraid1 format and will support Vraid0, Vraid1, Vraid5, and Vraid6 configurations.

- **Best practice for highest availability: Enhanced disk groups provide the highest levels of availability and metadata protection.**

RAID level influences availability

While Vraid5 provides availability and data protection features sufficient for most high-availability applications, some applications may require the additional availability and data-protection features of Vraid6 or Vraid1. Vraid6 or Vraid1 configurations can continue operation in failure scenarios where Vraid5 cannot. For example, a statistical model of the P6000 EVA shows that, for an equivalent usable capacity, Vraid1 provides over four times the data protection of Vraid5.[1]

This additional redundancy comes with additional cost caused by additional storage overhead. Nevertheless, some applications or file sets within an application warrant this additional protection.

[1] A statistical analysis is a prediction of behavior for a large population; it is not a guarantee of operation for a single unit. Any single unit may experience significantly different results.

If performance constraints do not allow a total Vraid6 configuration, consider using Vraid1 for critical files or data sets. For example:

- In database applications, select Vraid1 for log files.
- In snapshot and clone applications, select Vraid1 for active data sets and Vraid5 for snapshots and clones.

- **Best practice for highest availability: Vraid6 provides the highest levels of availability and data protection.**

Note:
The higher-availability and data-protection capabilities of Vraid1, Vraid5, or Vraid6 should not be considered a replacement for good backup and disaster-recovery processes. The best practices for business-critical applications always include frequent data backup to other near-line or offline media or devices.

Disk group organization influences availability and capacity utilization

Within a disk group, the P6000 EVA creates multiple sub-groups of disks called the RSS. Each RSS contains sufficient redundancy information to continue operation in the event of a disk failure within that RSS. The EVA can thus sustain multiple, simultaneous disk failures while not losing user data, as long as no more than one disk per RSS fails with Vraid1 and Vraid5, and no more than two disks per RSS fail with Vraid6. RSSs are created when a disk group is created, and additional sets are created as necessary when disks are added to the disk group. RSSs are created and managed by the EVA controllers, with no user intervention required, recommended, or supported.

The target size of each redundancy set is eight disks, with a minimum of six and a maximum of 11. As disks are added to a disk group, the RSS automatically expands until it reaches 12 disks. At that point, it splits into two sets of six disks each. As more disks are added, one set increases from six to eight (the target size); then the remaining set increases. After all disks have been added to a disk group, each RSS contains eight disks, with the possible exception of the last set, which contains between six and 11 disks. This is why it is a best practice to add disks in groups of eight. This is especially important when using Vraid6 when considering the various configuration related scenarios that can occur with RSS reorganization, including with the addition or removal of disks and potential disk failures.

- **Best practices to improve availability and capacity utilization:**
  - Keep the total number of disks in the disk group to a multiple of eight.
  - When creating a disk group, let the P6000 EVA choose which disks to place in the group.

Vraid0 influences availability

Unlike Vraid1, Vraid5, and Vraid6, Vraid0 provides no data redundancy. Vraid0 is best for applications where data protection is not a requirement. Because Vraid0 has no redundancy, data in Vraid0 requires less physical capacity, and performance is not affected by additional operations required to write redundancy information. Thus, Vraid0 provides the best performance for write-intensive workloads and the lowest cost of storage but the least availability. Use of Vraid0 may also impact the ability to upgrade disk drive firmware.

- **Vraid0 best practice to improve availability:**
  - Vraid0 is not advised for availability.
  - Vraid0 provides no disk failure protection.

Note:
For Vraid0, increasing the protection level does not increase the availability of the Vdisk. A single disk failure renders it inoperable, even when single or double protection is enabled.

Replacing a failed disk influences availability

Following the rules for shelf and disk organization is the best protection against potential data loss and loss of availability due to disk failure. However, when a disk fails, additional steps should be followed to reduce the risk of data loss or unavailability.

HP service engineers are trained on the proper EVA repair procedures and are alert to abnormal conditions that warrant additional steps to ensure continued operation. The best practice to maximize availability is to call for HP service. Customer self repair (CSR) is also an option for replacing a failed disk drive. If HP service is not an option or is unavailable, use the following rules.

When a disk fails, the EVA rebuilds the failed disk data through a process known as reconstruction. Reconstruction restores the disk group resiliency to protect against another disk failure. After reconstruction or after a new disk is added to a disk group, the EVA redistributes the data proportionately and reorganizes redundancy sets to the active disks.

- **Best practice to improve availability: Use the following procedure for disk replacement:**
  - Wait for the reconstruction to complete before removing the failed disks. This is signaled by an entry in the controller event log.
  - Use HP P6000 Command View to ungroup the disk. This ensures that the disk is not a member of a disk group.
  - Replace the failed disk. The new disk should preferably be inserted into the same slot as the failed disk. Ensure the disk addition policy is set to manual mode.
  - Verify and update the disk firmware version if needed.
  - Manually add the new disk into the original disk group.

Note:
CSR is also an option for replacing a failed disk drive. If a disk fails, order a replacement immediately and replace the disk as soon as possible using the following procedure.

Protection level influences availability

The protection level defines the number of disk failure—auto reconstruction cycles that the array can accomplish without replacement of a failed disk. Following a disk failure, the controller re-creates the missing data from the parity information. The data is still available after the disk failure, but it is not protected from another disk failure until the reconstruction operation completes.

For example, a “single” protection level provides continued operation in the event of two disk failures, assuming the reconstruction of the first failed disk completes before the second disk fails.

For Vraid1 and Vraid5, protection level “none” provides resilience to a single disk failure; whereas Vraid6 provides resilience to a dual disk failure; however, this is not a best practice configuration. **Vraid0 offers no protection from a disk failure.**

Conversely, the statistical availability of disks and the typical service time to replace a failed disk (MTTR—mean time to repair) indicate that a "double" protection level is unnecessary in all but the most conservative installations. A mitigating condition would be a service time (MTTR) that exceeds seven days. In that case, a protection level of double might be considered.

- **Best practice to improve availability: Use a single protection level.**

Note:
Protection level reserved capacity is not associated with the occupancy alarm setting. These are independent controls.

Number of disk groups influences availability

Although the EVA offers numerous levels of data protection and redundancy, a catastrophic failure (that is, multiple, simultaneous failures that exceed the architectural redundancy) can result in loss of a disk group. An example would be the failure of a second disk in an RSS before the reconstruction operation is complete. The probability of these events is low; however, installations requiring the highest levels of data availability may require creating multiple disk groups for independent failure domains (a failure in one domain does not affect the availability characteristics of the other domains). Multiple groups result in a slightly higher cost of ownership and potentially lower performance, but may be justified by the increased availability.

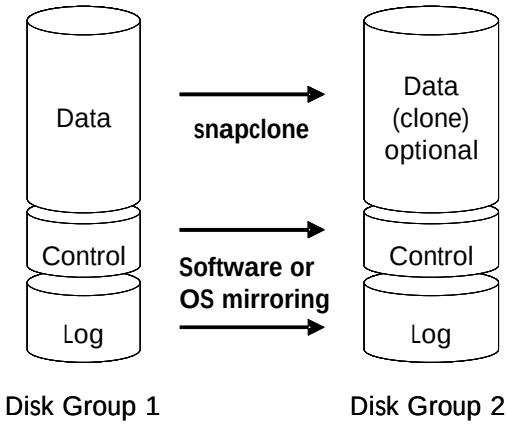
The strategy for multiple disk groups is to keep recovery data in a separate disk group from the source data. The typical use is to keep either a clone of the database or the database log files in a separate disk group from the application. If the primary disk group fails, the recovery disk group may remain available. Additional work is required to restore the application, but the recovery data is online, thus reducing the recovery time. Mirrorclones for instance, require a strategy of a fractured mirrorclone prior to the source LUN failure to provide access to the mirrorclone data, or snapshots of the attached mirrorclone can be utilized to aid in data restoration of a failed source LUN.

For two disk groups to prevent data loss, each disk group must contain sufficient independent information to reconstruct the entire application data set from the last backup. A practical example of this is a database that contains data files, configuration files, and log files. In this instance, placing the data files in one group and duplexing the log files and control files to both the data file disk group and another group ensure that loss of a single disk group does not prevent recovery of the data. (Duplexing or mirroring is a feature of some databases.)

An example is shown in Figure 1:

- Disk Group 1 contains data files, a copy of online redo logs, a copy of the control file, and an optional copy of archived logs (if supported by either the database or OS).
- Disk Group 2 contains a copy of online redo logs, a copy of the control file, the primary archive log directory, and an optional snapclone (or mirrorclone) of the data files for Disk Group 1.

Figure 1: Disk group example 1

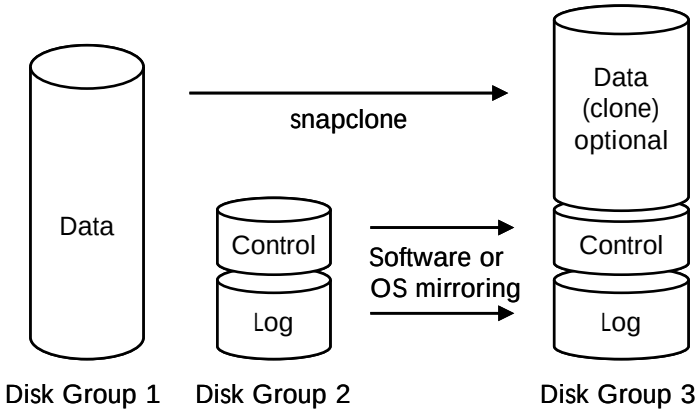


If Disk Group 2 does not contain the snapclone of the data files, the number of disks in Disk Group 2 should be determined by the sequential performance demand of the log workload. Typically, this results in more usable capacity than is required for the logs. In this case, choose Vraid1 for the log disks. Vraid1 offers the highest availability and, in this case, does not affect the cost of the solution.

A variation on this configuration is shown in Figure 2. In this example, two separate disk groups are used for the log and control files, and a third disk group is used for the data files. This configuration has a slightly higher cost but appeals to those looking for symmetry in the configuration. In this configuration:

- Disk Group 1 contains database data files.
- Disk Group 2 contains the database log, control file, and archived log files.
- Disk Group 3 contains a database log copy, control file copy, and archived log files copy (if supported).

Figure 2: Disk group example 2



Disk groups can be shared with multiple databases. It is not a best practice to create a separate disk group for each database.

- **Best practice to improve availability: For critical database applications, consider placing data files and recovery files in separate disk groups.**
- **Best practice to improve availability: Assign clones to a separate disk group.**

Note:
Creating multiple disk groups for redundancy and then using a volume manager to stripe data from a single application across both disk groups defeats the availability value of multiple disk groups. If, for cost or capacity reasons, multiple disk groups are not implemented, the next best practice is to store the database log, control file, and log archives in Vraid6 LUNs. Vraid6 provides greater protection from disk failures than Vraid1, Vraid5, or Vraid0.

Number of disks in disk groups influences availability

Disks assigned to disk groups in multiples of eight provide optimum placement of RSS disk sets. In this case, the controller can place the optimum number of disks in each RSS.

- **Best practice to improve availability: Size disk groups in multiples of eight disks.**

Capacity management and improved availability

Free space, or the capacity that is not allocated to a Vdisk, is used by the EVA controller for multiple purposes. Although the array is designed to operate fully allocated, functions like snapshot, dynamic LUN/RAID migration, thin provisioning, reconstruction, leveling, remote replication, and disk management either require additional free space or work more efficiently with it.

Three controls manage free space in the EVA: the protection level, the capacity occupancy alarm, and the capacity reported as available for Vdisk creation. Successful capacity planning requires understanding specific requirements for availability and cost, and setting the appropriate protection level, capacity occupancy alarm, and total Vdisk capacity.

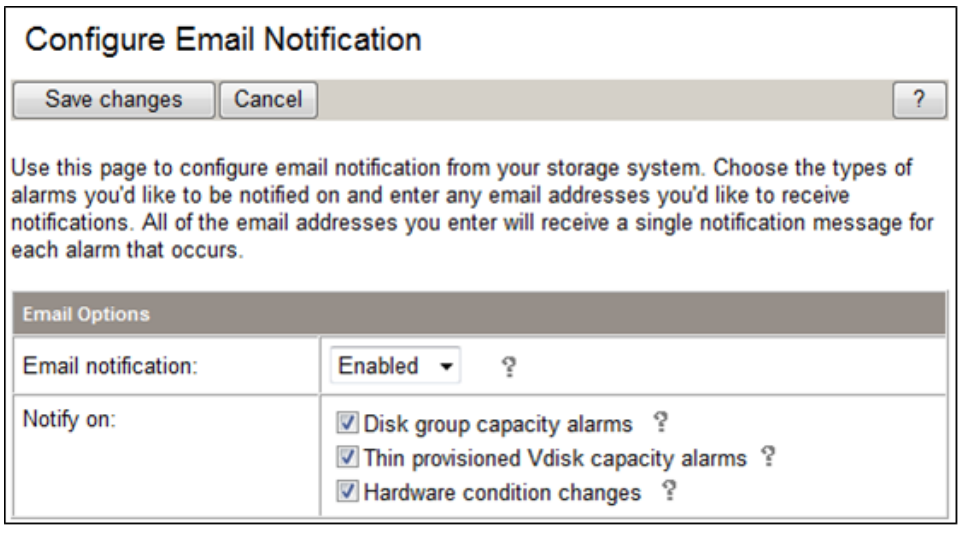
Set the protection level for the disk group. See the previous discussion of protection level and availability.

Additional reserved free space—as managed by the occupancy alarm and the total Vdisk capacity—affects leveling, remote replication, local replication, and proactive disk management. Figure 4 illustrates a typical disk group capacity allocation.

Occupancy alarm setting: The occupancy alarm is set for each disk group as a percentage of the raw capacity. Base the occupancy alarm setting on the unique installation requirements for proactive disk management, remote replication, and leveling.

Disk Group capacity alarm setting: HP P6000 EVA firmware XCS 10.00.00.00 introduces a warning level alarm on a disk group in addition to the critical level alarm from prior versions of XCS. This additional alarm level should be used to identify disk groups that are allocating space more rapidly than expected. This is an important notification mechanism for disk groups that contain demand–allocated snapshots, thin provisioned Vdisks, or both. The best practice is to use both alarm levels to provide early warning that the disk group capacity needs to be increased or Vdisks need to be moved to other disk groups (this capability is available in XCS 10.00.00.00 or later). In addition to SNMP traps and SMI-S indications, P6000 Command View offers email notification (see Figure 3 below). The best practice is to enable one or more notification mechanisms.

Figure 3: Configuring email notification of alarms



Proactive disk management: Proactive disk management (PDM) is a request by a customer or HP Services to ungroup a disk, or it is a predictive disk failure request by the EVA to ungroup a disk. In either case, the array migrates the contents of the disk to a free space before removing the disk from use. PDM can occur only if sufficient free space is available. PDM operation capacity is independent of protection level capacity. Customers who desire the highest levels of availability elect to reserve additional free space for PDM.

The capacity used for PDM is twice the largest disk in the disk group for each PDM event anticipated. Typical choices are none, one, or two events. The greater the disk count in a disk group, the greater the opportunity for a PDM event. See Protection level influences availability.

- **Best practice to enhance availability. Set the occupancy alarm to allow space for one or two PDM events per disk group.**

Remote replication: HP P6000 Continuous Access uses free space to allocate the DR group log (also known as the write history log). The DR group log is utilized when the remote link fails or is suspended. Until the link is reestablished, the EVA controller records changes locally in the DR group log. For free space management, plan for the additional disk capacity required for each DR group write history log. The size of the write history log is specified when the DR group is created.

Leveling and reconstruction: Leveling and reconstruction can be performed with a minimum of 5 GB of free space per disk group.

- **Best practice to improve availability.** Set the critical occupancy alarm to the capacity required for PDM plus 5 GB. This capacity is converted into a percentage of the raw capacity and then rounded to the next largest whole number. The pseudo-Excel formula would be as follows:

$$\text{Occupancy\_Alarm} = 100 - \text{CEILING}(\{\frac{\text{PDM\_capacity+5 GB}}{\text{total\_disk-group\_raw\_capacity}}\} * 100)$$

Where CEILING rounds the result in the parentheses up to the nearest integer.

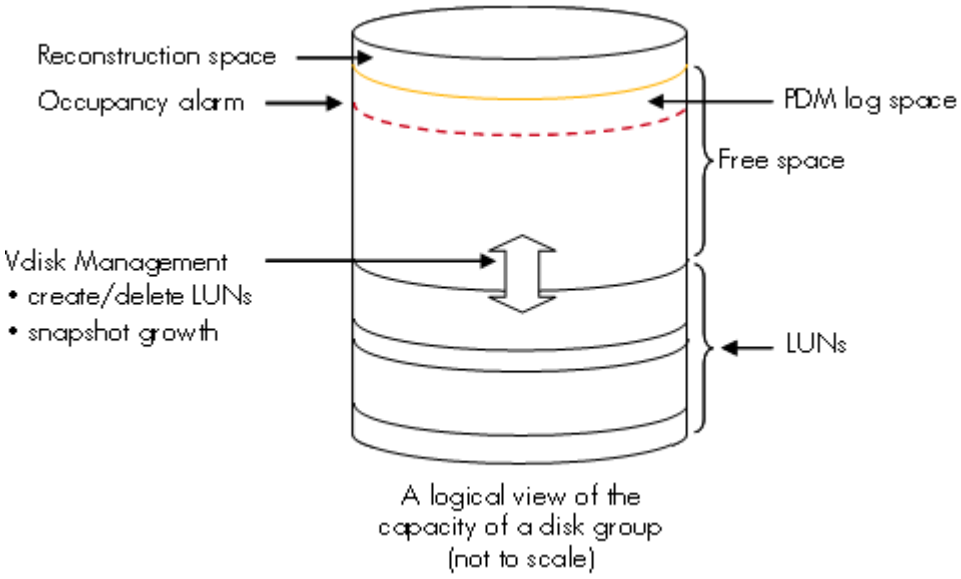
All capacities should be entered in GB. The result will be a number between 0 and 100, representing the percentage at which the Occupancy Alarm should be set.

- The following hypothetical example illustrates the calculation. Assume that the total disk group raw capacity is 40,000 GB and the PDM capacity is 1200 GB.

$$\begin{aligned}\text{Occupancy\_Alarm} &= 100 - \text{CEILING}(\{\frac{1200+5}{40000}\} * 100) \\ &= 100 - \text{CEILING}(\{0.0301\} * 100) \\ &= 100 - \text{CEILING}(3.01) \\ &= 100 - 4 \\ &= 96\end{aligned}$$

Thus the Occupancy Alarm should be set to 96 percent.

Figure 4: Disk group capacity allocation



Remaining free space: Remaining free space is managed by the creation of Vdisks, and is measured by the capacity available to create additional Vdisks. This free-space capacity is used by space-efficient snapshots and Thin Provisioning LUNs. It is critical that sufficient free space be available for the space-efficient snapshot copies and Thin Provisioning LUNs. If adequate free space is not available, all snapshot copies in the disk group become inoperative. (Space-efficient snapshot Vdisks become inoperative individually as each attempts to allocate additional free space. In practice, the effect is that all become inoperative together.) Fully allocated snapshot Vdisks and clone Vdisks will continue to be available. Thin Provisioning LUNs will enter an over commit state as writes are attempted to previously unallocated space. The behavior of over committed Thin Provisioning LUNs varies by OS, (see Thin Provisioning user guide for expected behavior).

Snapshots use copy-on-write technology. Copy-on-write occurs only when either the original Vdisk or the snapshot Vdisk is modified (a write); then an image of the associated blocks is duplicated into free space. For any given block, the data is copied only once. As snapshot copies diverge, the capacity available to create Vdisks decreases.

The actual capacity required for a space-efficient snapshot depends on the divergence of the original Vdisk and the snapshot, that is, how much of the data on the original disk changes after the snapshot is created. This value is unique for each application, but can range from 0 percent to 100 percent. The suggestion for the initial usage of snapshot (that is, when you do not know the actual physical capacity required by the snapshot) is to plan for (not allocate to Vdisks) 10 percent of the capacity of the parent Vdisks times the number of snapshots per parent Vdisk. For example, if you need to create two space-efficient snapshot Vdisks of a 500-GB Vdisk, you need to ensure that 100 GB (500 GB x 10 percent x 2) of usable capacity is available. Compute the usable capacity using the RAID level selected for the snapshot Vdisk.

- **Best practice to improve availability: Leave unallocated Vdisk capacity, the capacity available to create Vdisks, equal to the sum of the capacities required for all space-efficient snapshot copies within a disk group. Note that unallocated Vdisk space is also used by thin provisioning Vdisks. Monitor thresholds on the Vdisk and disk group to determine when to add more physical capacity to the disk group.**

- **Best practice to improve availability: Respond to a warning or critical occupancy alarm—evaluate what changed, replace failed disks, add disks, or reevaluate space-efficient snapshot or thin provisioning virtual disk usage. Extended operation of the array in an occupancy alarm condition is not a best practice.**

Fault monitoring to increase availability

The best protection from downtime is to avoid the failure. Good planning and early warning of problems can reduce or avoid many issues. For the EVA, Remote Support Package (RSP) is a free service that forwards EVA faults and warnings directly to HP Services through a secure virtual private network. HP can evaluate and diagnose problems remotely, possibly avoiding more serious issues. If an array requires service, RSP greatly increases the probability that our service engineer arrives onsite with the correct replacement parts to reducing the time to repair. If CSR is being used, RSP can check that the correct replacement part is provided.

If site security policies exclude direct communication to HP Services, RSP can be set up to report warnings to a customer contact. If RSP is not used, a custom automated alert process based on HP Web-based Enterprise Services (WEBES) or similar tool can be developed. It is critical that alarms from the EVA be carefully monitored.

- **Best practice to improve availability: Install and use RSP or equivalent tools to monitor and alert administrators to changes in EVA status.**

Integrity checking to increase availability

Exchange, Oracle, and other databases include tools to verify the integrity of the database. These tools check the database for consistency between records and data structures. Use these tools as part of the ongoing data center processes, as you would data backup and recovery testing. Proper use of these tools can detect database corruption early, when recovery options are still available.

- **Best practice to improve availability: Run database integrity checks as an ongoing data center process.**

Remote mirroring and availability

HP P6000 Continuous Access is an optional feature of the array that enables real-time remote mirroring. HP P6000 Continuous Access protects against catastrophic EVA or site failures by keeping simultaneous copies of selected LUNs at local and remote EVA sites. This feature can work standalone or in combination with system clustering software.

- **Best practice to provide the highest levels of data availability: Consider array-based HP P6000 Continuous Access or operating system-based replication software to provide real-time mirroring to a second EVA.**

HP P6000 Continuous Access and Vraid0 influence availability

When HP P6000 Continuous Access is used, LUNs can be logically associated into DR groups. DR groups allow a database to recover a remote copy with transaction integrity.

Two characteristics of remote mirroring are consistency and currency. Currency refers to the time difference between the remote and local copies. Consistency refers to the content difference between the mirrored LUNs and the local LUNs. A consistent remote LUN is either equal to the content of the local LUNs or equal to the content of the local LUNs at a past point in time. Synchronous HP P6000 Continuous Access provides mirror copies that are both current and consistent. Asynchronous operation provides mirror copies that are consistent, but may not be current (they may be slightly delayed). To accomplish asynchronous consistency, HP P6000 Continuous Access maintains the remote write ordering within each DR group. The order of the remote writes is identical to the order of the local writes. The writes may be delayed by the transmission distance, but the content of the LUNs in a DR group matches the current or a previous state of the associated local LUNs.

If a remote LUN becomes unavailable, HP P6000 Continuous Access cannot continue to write to the remaining LUNs in the DR group without losing DR group consistency. If the EVA continued to write to the remaining LUNs and the unavailable LUN became available, its contents would not be consistent with the other LUNs of the DR group, and a database recovery at this point would lose transaction integrity. Thus, the availability of a remote DR group is tied to the availability of the individual LUNs. If the remote DR group contains a Vraid0 LUN, mirroring to the entire DR group is tied to the availability of that Vraid0 LUN. While it is desirable to use Vraid0 as the target of a remote mirror to save costs, the solution must tolerate the resulting loss of availability.

- **Best practice to improve availability: Do not use Vraid0 as a target for HP P6000 Continuous Access mirror.**

System fault-tolerant clusters and availability

Many operating systems support optional fault-tolerant configurations. In these configurations, multiple servers and arrays are grouped together with appropriate software to enable continued application operation in the event of system component failure. These solutions span the replication continuum from simple local mirroring to complex disaster recovery solutions.

HP P6000 Continuous Access can be used for simple mirroring to increase data availability or as a component of a complete disaster-recovery solution. A complete disaster-recovery solution automatically (or manually) transfers the applications to the remaining functional systems in the DR group in the event of a site or component failure.

The HP P6000 EVA is supported in many disaster-recovery solutions. HP provides two disaster recovery products: Cluster Extensions for Windows® and Linux, and Metrocluster for HP-UX. For more details, contact HP or your operating system vendor. For applications requiring the highest availability, these system-level solutions should be considered.

- **Best practice to improve application availability: Consider disaster recovery solutions to provide continuous application operation.**

Best practices to enhance performance

Note:
Unless otherwise noted, the performance best practices in this document refer to configurations without HP P6000 Continuous Access. Additional HP P6000 Continuous Access performance and configuration information can be found in the HP P6000 Continuous Access implementation guide.

Experience shows that high performance and low cost generally have an inverse relationship. While this continues to be true, the HP P6000 EVA virtualization technology can significantly reduce the cost of high performance. This section outlines configuration options for optimizing performance and price-for-performance, although sometimes at the expense of cost and availability objectives.

Array performance management typically follows one of two strategies: contention management or workload distribution. Contention management is the act (or art) of isolating different performance demands to independent array resources (that is, disks and controllers). The classic example of this is assigning database table space and log files to separate disks. The logic is that removing the contention between the two workloads improves the overall system performance.

The other strategy is workload distribution. In this strategy, the performance demand is evenly distributed across the widest set of array resources. The logic for this strategy is to reduce possible queuing delays by using the most resources within the array.

Before storage devices with write caches and large stripe sets, like the EVA, contention management was a good strategy. However, with the EVA and its virtualization technology, workload distribution is the simplest technique to maximizing real-world performance. While you can still manage EVA performance using contention management, the potential for error in matching the demand to the resources decreases the likelihood of achieving the best performance.

This is a key concept and the basis for many of the best practices for EVA performance optimization. This section explores ways to obtain the best performance through workload distribution.

Enhancing performance raises the issue of demand versus capability. Additional array performance improvements have very little effect on application performance when the performance capabilities of the array exceed the demand from applications. An analogy would be tuning a car engine. There are numerous techniques to increase the horsepower output of a car engine. However, the increased power has little effect if the driver continues to request that the car travel at a slow speed. If the driver does not demand the additional power, the capabilities go unused.

The best results from performance tuning are achieved when the analysis considers the whole system, not just the array. However, short of a complete analysis, the easiest way to determine if an application could take advantage of array performance tuning is to study the queues in the I/O subsystem on the server and in the array. If there is little or no I/O queuing, additional array performance tuning is unlikely to improve application performance. The suggestions are still valuable, but if the array performance capabilities exceed the demand (as indicated by little or no queuing); the suggestions may yield only a modest gain.

Number of disks influences performance

In general, adding more disks provides better performance under typical workloads. However, there is diminishing return from additional disks as the combined performance capability of the disks approaches one or more of the performance limits of the controller. The performance limits of the controller depend on the characteristics of the application workload. For example, sequential workloads require fewer disks than random workloads to reach the point of diminishing return.

Where additional disks are needed only for capacity and the application does not require additional performance, increasing the number of disks in the array to the maximum that can be supported is a sensible solution. An example of this is when the demand is below the capability of the current EVA configuration. This can be determined by examining the disk queue depths. If there is little or no disk queuing delay, additional disks will not improve the performance of the array or application. For these situations, adding disks to the existing EVA provides the lowest-cost solutions.

- **Best practice to maximize single array performance: Fill the EVA with as many disk drives as possible.**

Number of disk groups influences performance

The number of disk groups has no effect on the performance capabilities of the EVA. An EVA can achieve full performance with a single disk group.

For typical workloads, an increased number of disk drives in a Vdisk improves the performance potential of the Vdisk. Because a Vdisk can exist only within a single disk group, it follows that having a single disk group maximizes the performance capability.

Large disk groups allow the workload to be distributed across many disks. This distribution improves total disk utilization and results in the most work being completed by the array. However, some application performance is measured by low response times for small, random I/Os, and not the total work completed. In this case, sequential workloads can interfere with the response time of the random component. For these environments, a separate disk group for the sequential workload can reduce the impact on the random I/O response time.

An alternative approach to maximize performance with multiple disk groups is operating system software (that is, a volume manager) that can stripe data across multiple LUNs on multiple disk groups. This provides a similar distribution of the workload to the disks as multiple LUNs on a single disk group. However, this solution provides lower capacity utilization than a single disk group.

- **Best practice to enhance performance: Configure as few disk groups as possible.**

Note:
Before considering a single disk group, see the discussion in this paper on the number of disk groups and availability.

Traditional Fibre Channel disk performance influences array performance

For applications that perform large-block sequential I/O, such as data warehousing and decision support, disk speed has little or no effect on the net performance of the EVA. Disk data sheets confirm that the average sustained large-block transfer rates are similar for both 10k rpm and 15k rpm disks of the same generation. Accordingly, large capacity 10k rpm and 15k rpm disks make the most sense for large-block sequential workloads.

For applications that issue small-block random I/O, such as interactive databases, file and print servers, and mail servers, higher speed disk drives offer a substantial performance advantage. Workloads such as these can see gains of 30 percent to 40 percent in the request rate when changing from 10k rpm to the equivalent number of 15k rpm disks.

Although it seems contradictory to use 10k rpm disks for better performance for small-block random I/O workloads, there are instances in which 10k rpm disks provide either better performance or better price-for-performance. The key is the relationship of the cost of the disk, the performance of the disk, and the quantity of disks. Because the performance gains from a 15k rpm disk ranges from 30 percent to 40 percent, if the 15k rpm disks are significantly more expensive than the 10k rpm disks, then it makes sense to purchase a larger number of 10k rpm disks.

The performance improvement with 10k rpm disks can be achieved only when the workload is striped across the disks. Unlike traditional arrays, the EVA automatically stripes the data across all the disks in the disk group, making this optimization easy to achieve.

- **Best practice to enhance performance: 15k rpm disks provide the highest performance.**
- **Best practice to enhance price-performance: For the equivalent cost of using 15k rpm disks, consider using more 10k rpm disks.**

Vraid level influences performance

Performance optimization is a complex activity; many workload characteristics and array configuration options can influence array performance. Without a complete workload analysis, array performance is difficult to predict. However, given this condition, in general, Vraid1 provides better performance characteristics over a wider range of workloads then Vraid5. However, Vraid5 can provide superior performance for some sequential-write workloads. The workloads that are candidates for Vraid5 contain a high percentage of sequential-write operations, and the write record size must be in multiples of 8K bytes. The larger the record size, the greater the Vraid5 advantage.

- **Best practice for Vraid level and performance: Vraid1 provides the best performance over the widest range of workloads; however, Vraid5 is better for some sequential-write workloads.**

Vraid0 influences performance

Vraid0 provides the best random write workload performance; however, **Vraid0 offers no protection from a disk failure**. When the risk is well understood, Vraid0 can improve overall application performance. As an example of the risk associated with Vraid0, a Vraid0 Vdisk using 16 physical disks can expect two or three data loss events during the five-year life of the array.

Some applications can rely on other means for data protection. Replicated or test databases or temporary storage used by a database are examples. However, avoiding data loss by external replication only provides protection for the data. Loss of a Vdisk can interrupt service and require manual steps to recover the Vdisk and resume application operation.

- **Best practice for Vraid0 and performance: Accepting the possible data and availability loss (Vraid0 provides no protection from disk failure), consider Vraid0 for non-critical storage needs only.**

Mixing disk performance influences array performance

The EVA is designed to support heterogeneous disk types in capacity or performance. Mixing drives of different speeds in the same disk group does not slow down access for all the disks to the speed of the slowest drive in the group. Each disk is independent.

Traditional array management suggests or requires the separation of heterogeneous disks into separate groups. Following this suggestion on the EVA would negate one of the most powerful performance features of the EVA: the ability to easily stripe across many disks to improve the performance of the array.

Although grouping drives by type and speed may seem easier to manage, it is actually difficult to balance the demand to individual disk groups. Errors in this balance can result in a disk group being under or over-utilized. Although the individual disks in a disk group can be slower, the ability to realize the aggregate performance of a single disk group is easier than when using two disk groups.

- **Best practice to enhance the performance of an array containing disks of different performance characteristics: Combine disks with different performance characteristics in the same disk group. Do not create separate disk groups to enhance performance.**

Mixing disk capacities influences performance and capacity utilization

The EVA supports disk groups consisting of disks with different capacities, with the exception that SAS Mid-line disks must be in their own respective disk group. Mixed-capacity disk groups are the recommended best practice for the optimization of the cost of storage.

Note:
For performance optimization and efficient capacity utilization, it is recommended that disk groups contain only disks of the same capacity.

The EVA stripes LUNs across all the disks in a disk group. The amount of LUN capacity allocated to each disk is a function of the disk capacity and Vraid level. When disks of different capacities exist in the same disk group, the larger disks have more LUN data. Since more data is allocated to the larger disks, they are more likely to be accessed. In a disk group with a few, larger-capacity disks, the larger disks become fully utilized (in a performance sense) before the smaller disks. Although the EVA does not limit the concurrency of any disk in the disk group, the workload itself may require an I/O to a larger disk to complete before issuing another command. Thus, the perceived performance of the whole disk group can be limited by the larger disks.

When the larger disks are included in a disk group with smaller disks, there is no control over the demand to the larger disks. However, if the disks are placed in a separate disk group, the storage/system administrator can **attempt** to control the demand to the disk groups to match the available performance of the disk group.

Separate disk groups require additional management to balance the workload demand so that it scales with the performance capability of each disk group. For a performance-focused environment, however, this is the best alternative.

- **Best practice to enhance performance: Use disks with equal capacity in a disk group.**
- **Best practice for efficient capacity utilization: Use disks with equal capacity in a disk group.**

- **Best practice to enhance performance with disks of unequal capacity: Create separate disk groups for each disk capacity and manage the demand to each disk group.**

Read cache management influences performance

Read caching can be set either at Vdisk creation or dynamically. This parameter affects both random-accesses read caching and sequential (prefetch) caching, although the algorithms and cache usage are different for these workloads.

Both algorithms come into play only when they will have a positive effect on performance. Random-access caching is enabled and disabled automatically as the I/O workload changes, while pre-fetch caching comes into play only when a sequential read stream is detected. Because of this dynamic response to changing I/O, cache efficiency is maintained at a high level, and there is no negative impact on either cache usage or performance when an I/O workload is "cache unfriendly."

Because there is no negative impact of leaving cache enabled and there is always a chance of a performance gain through caching, read cache should always be left enabled.

- **Best practice to enhance performance: Always enable read cache.**

Controller balancing influences array performance

The EVA can present LUNs simultaneously through both controllers. This allows the EVA to be compatible with popular dynamic load balancing and path management software, such as Windows Multipathing software, HP-UX PVLinks, and VERITAS DMP, as well as HP Secure Path (contact HP for the latest compatibility matrix). Either controller can own a LUN, but a LUN can be simultaneously accessed through either controller.

Even though LUNs are presented through both controllers, it is still important to balance the workload between controllers. This is accomplished by balancing the LUN ownership between the controllers. Although ownership should be distributed by workload demand, you can initially assign LUNs to controllers by capacity. Performance data from EVAPerf can help you improve the controller load balance.

The path through the controller that owns a LUN typically provides the best performance. If dynamic path management software is used, select the shortest-service-time option. If static path management is the only option, configure the path to use a port on the owning controller for each LUN.

When a controller fails, the LUNs owned by the failing controller are automatically switched to the remaining controller. The failover is automatic; however, the failback is not, unless the failover/failback option is used or the multipathing software drivers provide failback. After a failed controller is replaced, ensure that LUN ownership has been automatically or manually reassigned.

- **Best practice to enhance performance: Manually load balance LUN ownership to both controllers. If dynamic path management software is used, select the shortest-service-time option. Use EVAPerf to measure the load on each controller, and then redistribute LUN ownership as required.**
- **Best practice to enhance performance: Verify if the LUN ownership is reassigned after a failed controller has been repaired.**

Note:
Striping LUNs within a disk group on the same controller provides no additional performance value. The EVA automatically stripes each LUN across all disks in a disk group.

Note:
HP P6000 Continuous Access requires that all LUNs within a DR group be owned by the same controller. Load balancing is performed at the DR group level and not at the individual LUN. Additional configuration information can be found in the HP P6000 Continuous Access implementation guide.

LUN count influences performance

The EVA can achieve full performance with only a few LUNs per controller. However, the default queuing settings for some operating systems and host bus adapters (HBA) can constrain the performance of an individual LUN. In this case, additional LUNs or an increased OS/HBA default queue depth per LUN eliminates this constraint.

Pay particular attention to queue depth management for HP-UX. For HP-UX, the default queue depth is eight I/Os per LUN. This is insufficient for typical configurations that use only a few LUNs. You can find additional information on this subject in publications within the HP-UX community.

To enhance overall performance of arrays with clones, minimize the number of Vdisks that are issued with a clone request. In this case HP recommends creating the minimum number of Vdisks and modifying the host operating system and/or HBA queue depth setting to provide sufficient queue with the small number of LUNs.

In Microsoft Windows environments, attention to the Windows version and the HBA is required to understand the default I/O queue management operation.

Do not overlook the performance value of queue settings and LUN count. This is a common configuration error that can dramatically reduce performance.

- **Best practice to enhance performance: Follow operating system and application requirements for LUN count. If clones are used, create as few LUNs as possible and manage the operating system and Fibre Channel adapter queuing appropriately.**

Transfer size influences sequential performance

Applications such as data warehouse and business intelligence that have a high percentage of sequential write application can improve array performance by ensuring that the write transfer size is greater or equal to 32K and is a multiple of 8K (for example, 32K, 40K, 64K, 72K, 80K, … 128K). These transfer sizes simplify the cache management algorithms (which reduce the controller overhead to process a command) and reduce the total number of I/Os required to transfer a given amount of data. Storage systems typically choose these write transfer sizes, but operating systems, file systems, and databases also provide settings to manage the default transfer size.

- **Best practice to improve sequential write performance: Tune the write-transfer size to be a multiple of 8K and no greater than 128K.**

Snapshots and clones influence performance

The EVA can make local copies of selected Vdisks. There are three types of copies: snapclones, mirrorclones, and copy-on-write snapshots (either demand–allocated or fully allocated). All types are easy to use and integrate well into typical data center processes.

The simplicity of use of snapshot and clone operations masks the internal complexity and data movement required to execute the copy commands. There are three phases to the execution of internal Vdisk copies: the metadata management, the write cache flush, and the data movement. Metadata management is the work the controller needs to perform to create the internal data structures to manage the new Vdisks. Data movement is the operation of copying the data. Metadata management is similar for all local copy operations; however, data movement differs.

A snapclone or a mirrorclone makes a complete copy of an existing Vdisk. A snapclone is a point-in-time copy of the source disk at the creation of the snapclone. A mirrorclone is synchronized to the source and is continually updated as the source content changes. A mirrorclone is a point-in-time copy of the source disk at the time the relationship is fractured; while fractured changes to the source are tracked. When the mirrorclone is resynchronized to the source, only the changes that occurred during the fracture must be copied to the mirrorclone.

Clone creation places an additional workload on the disks of the target disk groups—the actual data copy. This workload competes with the external workload during the creation of the clone. The observed impact is an increased command response time and a decrease in the maximum I/O requests Per Second (IOPs) that the disk group can maintain. This performance impact continues until the Vdisk is completely copied. When the cloning operation completes, the performance impact ceases.

Dynamic LUN/RAID Migration uses the existing mirrorclone functionality with a new option “migrate”. This option is available from the P6000 Command View GUI and the Storage System Scripting Utility (SSSU). This best practice addresses a potential performance impact when using the SET MULTIMIRROR MIGRATE command in SSSU. The migrate option swaps the identities of a synchronized mirrorclone with its source virtual disk. The SET MULTIMIRROR MIGRATE command allows up to 28 mirrorclones to be sent to the array for migration. The potential observed impact would be an increase in I/O latency while the migrate operation is executed. This impact ceases when the migrate operation is completed, each migration occurs in less than a second. However if multiple migrate commands are executed concurrently the I/O latency of the final mirrorclone migrate operations can be multiple seconds. HP recommends for data access that is highly sensitive to changes in I/O latency the maximum number of mirrorclones targeted for migration be limited to 8. The user should also validate that all migrate operations have completed prior to issuing additional multimirror migrate commands. SSSU will accept additional SET MULTIMIRROR MIGRATE commands prior to the completion of the migrate operation for previous commands. The successful completion of the disk group or Vraid change can be verified via the Command View GUI or SSSU.

Snapshots take advantage of the HP P6000 EVA virtualization technology and copy only changes between the two virtual disks. This typically reduces the total data movement and associated performance impact relative to clones. However, there it does impact the performance. Snapshot uses copy-on-write technology, meaning that a data set is copied during the host write operation. A data set is copied only once. After it diverges from the original Vdisk, it is not copied again. Like clones, each dataset copied competes with the host workload. However, for many applications, less than 10 percent of the data in a Vdisk typically changes over the life of the snapshot, so when this 10 percent has been copied, the performance impact of the copy-on-write ceases. Another characteristic of typical workloads is that the performance impact exponentially decays over time as the Vdisks diverge. In other words, the performance impact is greater on a new snapshot than on an aged snapshot.

Pre-allocated containers can be used to perform the metadata management phase of snapshot or clone creation, before actually creating the snapshot or clone and initiating the write-cache flush and data movement phases. This allows the overhead for the metadata management (the creation of a LUN) to be incurred once (and reused) during low-demand periods. Using pre-allocated containers can greatly improve response times during snapshot or clone creation. The controller software version for the P6300 and P6500 arrays includes the additional container type "Demand-allocated" allowing the overhead for the metadata management to be incurred once for the demand-allocated snapshot creation.

Remember, the clone copy operation is independent of the workload; the copy operation is initiated by the clone request, whereas the snapshot is driven by the workload and by its design, and must compete with the workload resources (that is, the disks).

To make a consistent copy of a LUN, the data on the associated disks must be current before the internal commitment for a snapshot or clone. This requires that the write cache for a snapshot or clone LUN be flushed to the disks before the snapshot or clone commitment. This operation is automatic with the snapshot or clone command. However, the performance impact of a snapshot or clone operation can be reduced by transitioning the LUN to write-through mode (no write caching) before issuing the snapshot or clone copy command, and then re-enabling write caching after the snapshot or clone is initiated. The performance benefit is greatest when there are multiple snapshots or clones for a single LUN.

There is an additional step available to enhance the creation performance of either type of local business copy via the SSSU. The "prepare" command ensures that the original Vdisk and the container used for the creation of the snap shot, are both managed by the same controller.

Snapshots affect the time required for the array to perform a controller failover, should it become necessary. Failover time is impacted by the size of the source Vdisk and the number of snapshots of the source. To minimize the controller failover time and host I/O interruption, the size of the source Vdisk multiplied by the number of source snapshots should not exceed 80 TB.

(Size of source virtual disk) X (Number of source snapshots) ≤ 80 TB

- **Best practice for snapshot and clone performance:**
  - Use pre-allocated containers, fully-allocated or demand-allocated for snapshots and fully-allocated for clones. Transition source LUN to write-through cache mode before snapshot or clone initiation, and re-enable write caching after the snapshot or clone is initiated.
  - Use the SSSU "prepare" command to ensure the Vdisk and the container are managed by the same controller, (command only available in the SSSU interface).
  - Create the snapshot within 60 seconds of the prepare command execution for maximum performance.
  - Create and delete snapshots and clones during low-demand periods, or size the array to meet performance demands during snapshot or clone activities.
  - Use the SSSU "multisnap" command to limit the total capacity of Vdisks being snapped to less than 32 TB.

- **Best practice for clone performance:**
  - Keep virtual disks as small as possible.
  - Minimize the concurrent clone operations (use fewer Vdisks). Organize clone operations into consistency groups of Vdisks, and then clone consistency groups sequentially.

- **Best Practice for SET MIRRORCLONE MIGRATE performance:**
  - Minimize concurrent mirrorclone migrate operations in the SSSU multimirror migrate command to 8 mirrorclones
- **Best practice for snapshot performance:**
  - Minimize the number of Vdisks with active snapshot copies. Use fewer Vdisks (it is better to have a few large Vdisks than many small Vdisks).
  - Minimize the number of snapshot copies for a Vdisk. Do not keep extra snapshot copies without reason or plan for their use.
  - Minimize the life of a snapshot copy. If snapshot copies are used for backup, consider deleting the snapshot Vdisk at the completion of the copy to tape.
  - Delete snapshot Vdisks in order of age; oldest first.
  - Minimize the controller failover time and host I/O interruption, by ensuring that the size of the source Vdisk multiplied by the number of its snapshots does not exceed 80 TB.

Space-efficient snapshots and TP LUNs use free space (capacity not reserved for normal or clone Vdisks) to store data. All space-efficient snapshot Vdisks in a disk group become inoperative when any space-efficient snapshot Vdisk in that disk group is denied a request to use additional free space. Always monitor free space. If the availability of space-efficient snapshot or thin provisioned Vdisks is critical for the whole application availability, then overestimating the requirements for free space may be warranted.

In addition to the normal LUN divergence consumption of free space, a disk failure and the subsequent reconstruction can also compete for free space. After a reconstruction, the reserved space requirements for the protection level can cause the existing snapshot Vdisks to exceed the available free space and so can cause the snapshot Vdisks to become inoperative. A thinly provisioned Vdisk would transition to over commit only on a write request (to unallocated capacity) after all available capacity has been allocated. See the best practice for capacity management and improved availability to avoid this condition.

HP P6000 Continuous Access and snapshots influence performance

Improvements in overall system performance can be realized when making snapshot copies of the remote target of an HP P6000 Continuous Access pair by temporarily suspending its DR group at the initiation of the remote snapshot copy.

The performance impact of an improperly sized (in a performance sense) snapshot-copy LUN can cause the P6000 Continuous Access to suspend replication. System and application performance and availability can be improved by making this possible condition a planned event rather than an unplanned disruption.

- **Best practice to enhance performance when making snapshots of remote Vdisks in DR groups: Suspend the DR Group, create the snapshots, and then resume the DR group.**

Miscellaneous management best practices

Increasing capacity of the array

To reduce false indications of excessive errors, insert multiple disks carefully and slowly, pausing between disks. This cautiousness allows the initial bus interruption from the insertion and the disk power-on communication with the controller to occur without the potential interruption from other disks. In addition, this process sequences leveling so that it does not start until all the new disks are ready.

Although the array supports replacing existing smaller disks with larger disks, this process is time consuming and disruptive and can result in a non-optimum configuration. Do this only if the option to build new disk groups and move existing data to the new disks is unavailable.

- **Best practice to improve availability when adding disks to an array:**

  - Set the add disk option to manual.
  - Add disks one at a time, waiting a minimum of 60 seconds between disks.
  - Distribute disks vertically and as evenly as possible to all the shelves.
  - Unless otherwise indicated, add new disks to existing disk groups using the HP Storage System Scripting Utility add multiple disks command.
  - Add disks in groups of eight.
  - For growing existing applications, if the operating system supports Vdisk growth, increase Vdisk size. Otherwise, use a software volume manager to add new Vdisks to applications.

Disk groups and data security

Disk groups are self-contained components of the array; that is, the storage resources required for a disk group are contained completely within each disk group. A given disk group can be accessed only through LUNs created from that disk group. Also, a LUN can contain capacity from only one disk group.

Given these characteristics, applications or data centers that require data isolation for security objectives can accomplish these objectives by assigning unique security domains to separate disk groups.

These characteristics also make disk groups useful for tracking and allocating assets to specific groups within an organization.

- **Best practices for data security: Assign application and servers to separate disk groups. Use selective LUN presentation to limit access to approved servers.**

Note:
Multiple disk groups increase the cost of the storage and may reduce the performance capability of the array. See the sections on disk group usage.

Best practice folklore, urban legends, myths, and old best practices

Urban legend: Pre-filling new LUNs improves performance

There are two phases to the creation of a non-snapshot LUN on an HP P6000 EVA. The first phase creates the metadata data structures. The second phase writes zeros to the associated sectors on the disks. Access to the new LUN is restricted during the first phase, but access is allowed during the second phase. During this second phase, host I/Os compete with the zeroing operation for the disks and controller, and performance is impacted.

When the zeroing completes, the array is capable of delivering normal performance for that LUN. The duration of the zeroing operation depends on the size and Vraid level of the LUNs, the number of disks in the disk group, and the host demand. In a test case, using 168 36-GB disks to create 64 46-GB LUNs, zeroing required 30 minutes to complete and there was no other load on the EVA.

Zeroing is a background operation, and the time required to complete zeroing increases when other workloads are present on the array. The Vdisk is fully accessible during this time; however, the performance of the Vdisk during zeroing does not represent future performance.

For non-snapshot Vdisks, the EVA always maps data on the disks in its logical order. Unlike other virtual arrays for which the layout is dynamic and based on the write order, the EVA data structure is predictable and repeatable.

Given the new LUN-zeroing operation and the predictable data layout, there is no reason (with the exception of benchmarking) to pre-allocate data by writing zeros to the Vdisk on the EVA.

Note:
Pre-filling a new LUN before a performance benchmark allows the internal zeroing operation to complete before the benchmark begins.

Most misunderstood best practice: 5 GB of free space is sufficient for array optimization in all configurations, or 90 percent/95 percent LUN allocation is a good rule-of-thumb for free space.

These rules are over simplifications and are correct for only some configurations. Five GB is sufficient only if no demand-allocated snapshots or thin provisioned Vdisks exist and some availability tradeoffs are accepted (proactive disk management events). Ninety percent is frequently more than what is actually required for optimum operation, thus unnecessarily increasing the cost of storage. Be sure to read and follow the free space-management best practices.

Summary

All of the preceding recommendations can be summarized in a table. This not only makes it relatively easy to choose between the various possibilities, but it also highlights the fact that many best practice recommendations contradict each other. In many cases, there is no single correct choice because the best choice depends on the goal's cost, availability, or performance. In some cases, a choice has no impact.

Table 2: Summary

	Cost	Discussion	Performance
Mixed disk capacities in a disk group	Yes	—	No
Number of disk groups[2]	1	>1	As few as possible
Number of disks in a group	Maximum	Multiple of 8	Maximum
Total number of disks	Maximum	Multiple of 8	Maximum
Higher performance disks	No	—	Probability
Mixed disk speeds in a disk group	Yes	—	Acceptable
Protection level	0	1 or 2	1
LUN count[3]	—	—	—
Read cache	—	—	Enabled
LUN balancing	—	—	Yes

[2] Consult application or operating system best practices for minimum number of disk groups.
[3] Check operating system requirements for any special queue depth-management requirements.

Glossary

access density	A unit of performance measurement, expressed in I/Os per second per unit of storage. Example, I/Os per second per GB.
clones	A term used to refer collectively to both snapclones and mirrorclones.
data availability	The ability to have access to data.
data protection	The ability to protect the data from loss or corruption.
disk group	A collection of disks within the array. Vdisks (LUNs) are created from a single disk group. Data from a single Vdisk is striped across all disks in the disk group.
Dynamic capacity management (DC-Management)	An EVA feature that enables the automatic resizing of a Vdisk.
free space	Capacity within a disk group not allocated to a LUN.
leveling	The process of redistributing data to existing disks. Adding, removing, or reconstruction initiates leveling.
LUN	Logical unit number. An addressable storage collection. Also known as a Vdisk.
occupancy	The ratio of the used physical capacity to the total available physical capacity of a disk group.
physical space	The total raw capacity of the number of disks installed in the EVA. This capacity includes protected space and spare capacity (usable capacity).
protection level	The setting that defines the reserved space used to rebuild the data after a disk failure. A protection level of none, single, or double is assigned for each disk group at the time the disk group is created.
protection space	The capacity that is reserved based on the protection level.
reconstruction, rebuild, sparing	Terms used to describe the process of recreating the data on a failed disk. The data is recreated on spare disk space.
reserved space	Same as protection space.
RSS	Redundancy Storage Set. A group of disks within a disk group that contain a complete set of parity information.
TP	Thin Provisioning
usable capacity	The capacity that is usable for customer data under normal operation.
virtual disk	A LUN. The logical entity created from a disk group and made available to the server and application.
workload	The characteristics of the host I/Os presented to the array. Described by transfer size, read/write ratios, randomness, arrival rate, and other metrics.

To simplify management with HP P6000 EVA and to reduce time and money, visit www.hp.com/go/P6000

HP-UX Release 10.20 and later and HP-UX Release 11.00 and later (in both 32 and 64-bit configurations) on all HP 9000 computers are Open Group UNIX 95 branded products.

4AA3-2641ENW, Created June 2011